Introduction: The added value of diachronic treebanks for historical linguistics research

by Hanne Martine Eckhoff (University of Oxford, hanne.eckhoff@mod-langs.ox.ac.uk), Silvia Luraghi (University of Pavia, silvia.luraghi@unipv.it) and Marco Passarotti (Catholic University of the Sacred Heart, marco.passarotti@unicatt.it)

1 Ancient languages and digital sources

Over the last few decades, the wide diffusion of digital technology and the growing ease of transferring information via the Internet have provided scholars with an enormous amount of textual data at hand. The vastly increased availability of primary sources has radically changed the everyday life of scholars in the humanities, who are now enabled to access, query, and process a wealth of empirical evidence like never before.

The development also encompasses ancient languages. The first aim in the eighties and the nineties was to digitize textual data and make them available on CD-ROM and online. Later, the need for linguistic annotation gave rise to projects aimed at building corpora enhanced with increasingly complex layers of metalinguistic information, such as part-of-speech tagging and syntactic annotation, opening the field to precise queries for particular linguistic phenomena. We are now at a stage at which several of these syntactically annotated corpora, or treebanks, have reached a mature state, providing representative selections of texts for several diachronic stages of the language in question. These new linguistic resources allow a new approach to

diachronic studies of syntactic phenomena where scholars previously had to content themselves with data work on a much smaller scale.

This volume collects a set of papers that report research on various diachronic matters supported by evidence provided by diachronic treebanks for different languages, i.e. treebanks that provide data for the language across several historical stages. We show that diachronic treebanks not only allow us to shed new light on vexed topics in the literature, but can also provide considerable methodological advances – greater transparency and better ways of exploiting the frequently problematic source material.

2 What is a treebank?

In linguistics and philology, the term "corpus" has traditionally been used simply to denote a set of texts used to explore some linguistic phenomenon. Many types of digital text resources are now also referred to as "corpora". McEnery et al. (2006) simply define a corpus as "a body of naturally occurring language", while Sinclair (2005) gives a much stricter definition: "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". However, even under the stricter definition, there is no requirement that the corpus should have linguistic annotation of any kind. Thus there is vast variation in the amount of work that has gone into corpora and the usefulness of the resource for linguists researching particular phenomena. A corpus may be anything from a digitized, machine-readable text collection that only allows queries for text strings, to a sophisticated, multi-

layered text resource with several types of linguistic markup, queriable by a dedicated query engine. In this volume, we concern ourselves with one of the most work-intensive corpus types of all: the treebank.

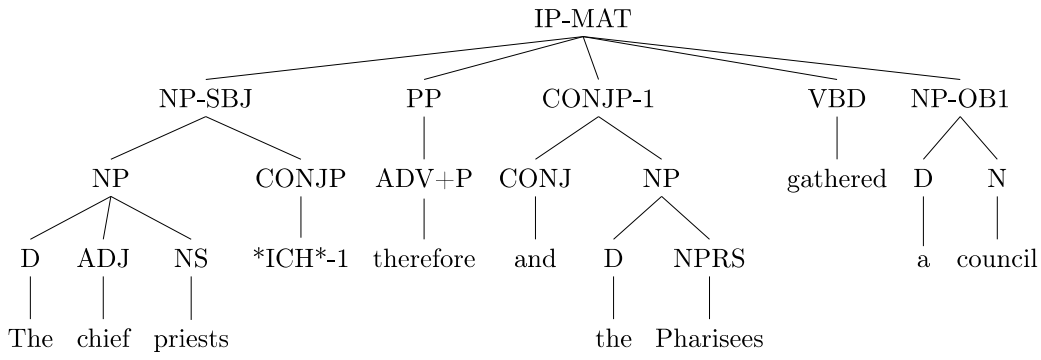
A treebank is a text corpus with exhaustive syntactic annotation, typically applied on top of lemmatization, part-of-speech tagging and morphological annotation. Each of these annotation layers add to the precision of queries. Lemmatization allows for queries for all word forms subsumed under a single lemma, with no need for use of regular expressions. Part-of-speech and morphological tags allow for queries for specific combinations of linguistic features at word level, without obligatory reference to the word form. Syntactic tagging makes it possible to search for groups of words that are syntactically related, regardless of whether they are adjacent to each other or not. Since syntactic queries will mostly be multi-word queries, and syntactic queries are typically combined with features from other layers, they can quickly become quite complex, and will either require a good query engine or that the users master a query language. However, given such facilities, a treebank offers queries of unparalleled precision: if the annotation is good enough, it is possible to make queries almost entirely free of noise. For example, one may find all infinitives with preverbal pronominal direct objects in a single query.

Although some treebanks are annotated in accordance with the formalism of a particular syntactic framework, most strive to be relatively theory-neutral. There are two major groups of annotation schemes: phrase-structure-based schemes and dependency-based schemes. The first major treebank to be released, the Penn Treebank (1989–1996, Taylor, Mitchell & Santorini 2003), used a (simplified) phrase-structure scheme, which is continued in the many Penn daughter treebanks. There are also numerous dependency treebanks, many inspired by the groundbreaking

Prague Dependency Treebank (Bejček et al. 2013). Existing dependency treebanks employ a number of competing annotation schemes. As a reaction to this, the Universal Dependencies[1] (UD) initiative has developed a universal consensus-based scheme, and works to convert as many treebanks as possible into the UD scheme.

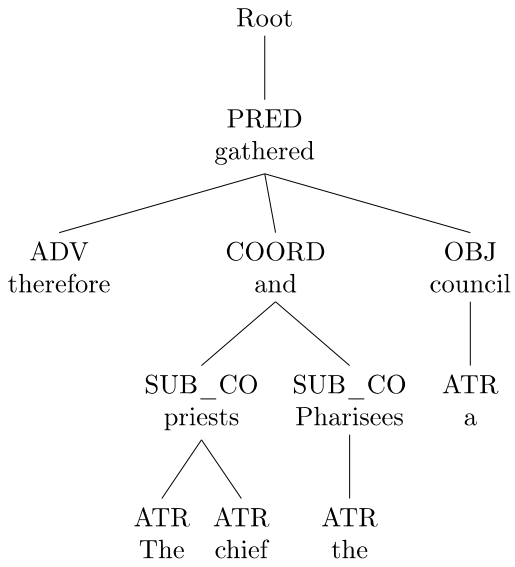
The two main treebank styles are based on two syntactic notions that both clearly have some psychological reality. Phrase-structure treebanks are based on the idea that words are organized into groups (constituents) with certain properties, for example, that the whole constituent can be substituted by a pro-word, and will normally move together. Dependency treebanks, on the other hand, are based on the idea that every word in a sentence will have one and only one syntactic head. As a brief illustration of the differences between these two main treebank styles, consider the two syntactic trees below. Tree (1) is the original Penn-style analysis of the opening of John 11.47 (American Standard Version, see Taylor & Pintzuk this volume, p. XX). Tree (2) is the same passage analyzed in the Prague Dependency Treebank format. Both analyses are given in tree format for ease of comparison. Figure (3) presents the dependency analysis in linear order.

(1)

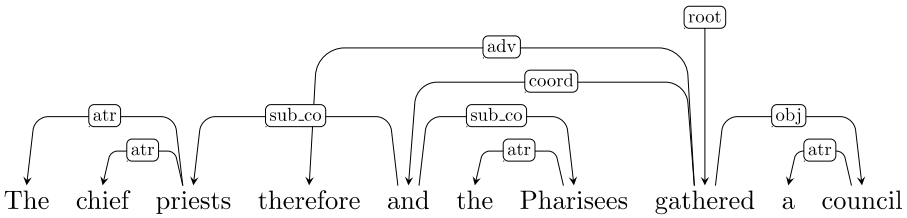


[1] http://universaldependencies.org/

(2)



(3)



We see that the phrase-structure analysis in the Penn-style tree is fairly flat, which brings the two analyses closer than they might have been if the Penn scheme had been binary-branching. The most striking difference in this example is that the Penn analysis cannot have crossing branches, and therefore deals with split coordination (the topic of Taylor and Pintzuk's paper) with a trace (the *ICH*-1). The index of the trace is then picked up again in the CONJP-1, the second part of the coordination which is represented in its linear place in the sentence. In the dependency analysis, the fact that the coordination is split is not represented at all, and can only be retrieved by

combining the dependency analysis with word order information stored in a different layer (visualized in (3)). However, this analysis is computationally simpler, since every node in the tree corresponds to a lexical item.

3 Historical corpora and treebanks

Historical linguistics has always relied on corpora. This observation is captured by the German use of the term *Korpussprachen* to refer to dead languages. Indeed, with most dead languages, all we have is a more or less extended corpus of written texts. This constitutes a limitation (one cannot check if something that is not in the corpus is not there because it is ungrammatical, or just by accident), but also makes it possible for linguists working on such languages to base their assumptions on all attested forms. Extended corpora, even if finite, often exceed the linguist's possibility to check all occurrences: for this reason, the introduction of digitized corpora has been a welcome addition to historical linguistics research, in much the same way as in research on spoken languages. Parsed corpora have the further advantage of adding information at various levels of linguistic analysis through the addition of metadata. Among these, treebanks have become an increasingly useful resource for data-driven study of linguistic structures at various levels (Freddi & Luraghi 2013).

Diachronic treebanks are closely related to existing synchronic treebanks, and generally use their annotation schemes or schemes inspired by such schemes. For example, there are several diachronic treebanks using the Penn phrase-structure format, such as the York-Toronto-Helsinki Corpus (http://www-users.york.ac.uk/~lang22/YcoeHome1.htm), the Penn Corpora of Historical English

(https://www.ling.upenn.edu/hist-corpora/) for English (Taylor, Warner, Pintzuk & Beths 2003; Kroch & Taylor 2000; Kroch, Santorini & Delfs 2004), and the French diachronic treebank "Modéliser le changement : les voies du français" (MCVF, Martineau 2008, http://www.voies.uottawa.ca/corpus_pg_en.html), all represented in this volume. There are also several diachronic dependency treebanks. Some of them directly use the Prague Dependency Treebank format, for example the Perseus Latin and Ancient Greek Dependency Treebanks (http://nlp.perseus.tufts.edu/syntax/treebank/) (Bamman & Crane 2011), and the *Index Thomisticus* Treebank (http://itreebank.marginalia.it) (Passarotti 2011). Other diachronic dependency treebanks, such as the PROIEL family of treebanks (Haug & Jøhndal 2008, Eckhoff & Berdičevskis 2015, Eckhoff et al. 2017) and the Syntactic Reference Corpus of Medieval French (SRCMF, Stein & Prévost 2013), have developed annotation schemes of their own, inspired by more classic schemes, but with modifications intended to increase expressivity.

Diachronic treebanks must also take part in projects of treebank standardization such as the Universal Dependencies initiative, and many are already doing so, for instance, both PROIEL and Perseus have converted their diachronic Ancient Greek treebanks to Universal Dependencies format and released them.[2] This will enable users to run comparative diachronic analyses on a multilingual scale, and will strengthen the essential role that treebanks will increasingly play in the near future, in historical linguistics as well as in general linguistics.

Nonetheless, diachronic treebanks, like synchronic treebanks for historical/ancient languages, show a number of peculiar features that make them

[2] http://universaldependencies.org/treebanks/grc_proiel/index.html, http://universaldependencies.org/treebanks/grc_perseus/index.html

different from those for modern languages. These features concern both the building and the use of such linguistic resources.

On the building side, diachronic treebanks raise specific issues concerning the selection, nature, and size of the source data. While selectional criteria are a core issue in the building of synchronic treebanks for modern languages, diachronic treebanks are obliged to exploit to the maximum the few (and thus most precious) texts that remain from the past. This is an obvious consequence of data availability: while the bulk of available data for modern languages increases on a daily basis, corpora for ancient languages are closed, with no (or very few) new additions. Corpora for modern languages strive to be as representative as possible of a language or of one specific variety of it, and rely on a huge, in principle open-ended material. Diachronic treebanks, instead, tend to include full texts of single authors, or, often, all available texts, in order to compensate for the limited size of the available data.

Diachronic treebanks are usually smaller than those for modern languages, but the limited amount of available texts is typically not the only reason. Most efforts to build diachronic treebanks are quite recent, and the annotation work is typically more difficult and time-consuming than annotation of texts in modern languages. Modern treebanks can often balance shallow annotation with the availability of huge masses of data, as shown by the recent and already widespread use of deep learning techniques such as word embedding. Word embedding is a collective name for techniques that aim to quantify and categorize semantic similarities between linguistic items on the basis of their co-occurrence patterns in large datasets, converting their distributional profiles to vectors of numbers. Such techniques are extremely data-intensive and therefore normally out of reach for historical texts: there is not enough training data, nor is there enough data to afford the noise produced by automatic

shallow annotation. To be of value, the annotation must be at least manually corrected.

There are also complications brought about by editorial issues connected with historical texts. Because of the way in which they are collected, texts of treebanks for modern languages usually do not raise specific editorial issues, as they normally come in a single version. When dealing with historical texts, there can be several different versions of the same text, and choosing the critical edition to record in the corpus is a substantial part of the selection work, which also concerns the question whether or not to include apparatus information (such as variants) in diachronic treebanks.

A further difference lies in the nature of the source data. In most cases, texts included in treebanks for modern languages are digital in the outset, while those in diachronic treebanks result from digitization processes. Especially when digitization results from the application of OCR techniques, this can result in errors, which need to be found and corrected.

On the use side, treebanks including diachronic data differ from those for modern languages primarily through a different attitude to the texts themselves. Diachronic treebanks tend to include literary, historical, philosophical or documentary texts. Treebanks for modern languages, on the other hand, are often based on texts from newspapers. As a consequence, users of data from modern treebanks are interested in exploiting such resources as providers of empirical evidence either in support of general tendencies of a language or for different purposes in the area of natural language processing (such as, for instance, inducing grammars and training stochastic parsers). Users of diachronic treebanks, on the other hand, are in most cases interested in the texts themselves. This makes the results of the analyses run on such treebanks more bound to the specific kind of data provided by the resource in

question and, thus, less portable in terms of linguistic generalization. But such an interest in the very empirical evidence provided by diachronic treebanks also increases the value itself of the distinctive features of the texts included there, and supports research focused on specific linguistic aspects in a specific period of time, which is a major issue in diachronic studies. In this volume, Timo Korkiakangas' paper exemplifies work that exploits data from quite peculiar Latin texts (Medieval charters), whose results cannot be generalized on the entire Latin language but concern one of its specific instantiations.

4 Historical treebanks in use

So far only a small number of scholars in historical linguistics use treebanks in their everyday work, as demonstrated, among other things, by the limited number of papers describing treebank-based research published in this journal. This is partly due to the still low availability of representative diachronic treebanks, but partly also to the fact that many historical linguists are simply not informed about the very existence of such resources. Treebank providers must therefore improve the dissemination of their products. As a matter of fact, several papers in this volume are authored by scholars presenting research that makes use of treebank data that they have directly contributed to building, and it is one of our aims to foster broader use of treebanks in historical linguistics.

Diachronic treebanks allow data extraction aimed to assess the scope and effects of diachronic developments, managing a large amount of data and retrieving information whose relevance can then be evaluated through statistical methods. In this

special issue we especially emphasize the point made by Haug (2015): the use of treebanks in historical linguistics allows us to publish research that is truly replicable.

It is a well-known fact that publications in historical linguistics that deal with the same topic and use the same text sources may still reach very different conclusions and report very different statistics for the phenomenon under scrutiny. As an example, Haug (2015) cites word order statistics reported by a number of researchers for the Gospel of Luke and Acts (which were written by the same author). The statistics differ to an alarming extent, and the reason is that the researchers have used different selection criteria and made different theoretical assumptions, but have not necessarily made these criteria and assumptions explicit. This means that their results are impossible to compare and replicate. Research on several of the topics covered in this volume also suffers from such problems, for instance the literature on the rise of *po* delimitatives discussed by Eckhoff. Indeed, by extracting her data from the Tromsø Old Russian and OCS Treebank, Eckhoff shows that in the OCS and Old East Slavic data sets, the *po* delimitative is not as marginal as commonly claimed in the literature.

The use of treebanks can help to remedy this situation in two ways. First, a good treebank is annotated consistently using an annotation scheme that is founded on broad consensus in the linguistic community, such as the simplified phrase structures in the Penn annotation scheme or the notion of syntactic dependency in the various dependency grammar schemes. Such treebanks are not created for the aims of one specific line of research, and exploiting the empirical evidence provided by such (publicly available) linguistic resources prevents the vicious circle of creating a corpus with the specific aim of studying a single linguistic phenomenon (Sinclair

2005). Thus, simply using a good treebank will make a number of basic theoretical assumptions explicit.

Second, to retrieve data from a treebank requires a clear and explicit query expression which lists the selection criteria. Ideally, the query criteria should therefore be published with the research results, and so should the full data set. Any additional annotations and classifications that do not come directly from the treebank should be made explicit in the data set. Online data repositories offering persistent identifiers are the best way to publish such data sets. We would like this volume to serve as an example in this respect: all of the papers include the queries have been run to collect the empirical evidence used to support the authors' conclusions, either in appendices or in online repositiories, and several also provide full data sets and scripts used to process them. Given that the audience of *Diachronica* is not made up of computer scientists and computational linguists, the authors have also strived to explain the details of their queries (according to the specific query language they used) and analyses, and provide readers with the opportunity of running them themselves.

5 Aims and scope of this volume

This volume presents a series of studies that demonstrate the potential of a number of mature diachronic treebanks that are now available, represented by the following: For English, we use the York-Toronto-Helsinki corpus[3] and the Penn Corpora of Historical English[4] (Taylor, Warner, Pintzuk & Beths 2003; Kroch & Taylor 2000;

[3] http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

[4] https://www.ling.upenn.edu/hist-corpora/

Kroch, Santorini & Delfs 2004). For French, we use the MCVF corpus[5] (Martineau 2008). For Russian and Old Church Slavonic, we use the Tromsø Old Russian and OCS Treebank[6] (Eckhoff & Berdicevskis 2015), as well as PROIEL[7] (Haug & Jøhndal 2008), which is also used for Latin and Ancient Greek. For Latin and Ancient Greek, we also use the Late Latin Charter Treebank (Korkiakangas & Passarotti 2011), which was built along the lines of the Perseus Latin and Ancient Greek Dependency Treebanks[8] (Bamman & Crane 2011) and the *Index Thomisticus* Treebank[9] for Latin (Passarotti 2011).

Our aim is to demonstrate the multiple ways in which diachronic treebank data may be used to advance historical linguistics. Treebank data may not only shed new light on vexed topics in the literature, but also be the foundation of considerable methodological advances. Provided that they are built in a way that takes into consideration both the peculiarities of the text material and the overall standards in the treebank community, they may provide much-needed transparency and replicability to studies in historical linguistics. They may also be used to develop techniques that can to some extent compensate for the inherent problems of the historical text sources, which tend to be archaic, skewed to certain genres and often sparse. This volume contains papers that touch on all of these topics.

5.1 Old topics, new methods

[5] http://www.voies.uottawa.ca/corpus_pg_en.html

[6] https://nestor.uit.no, http://torottreebank.github.io/

[7] https://proiel.github.io/

[8] http://nlp.perseus.tufts.edu/syntax/treebank/

[9] http://itreebank.marginalia.it

The bulk of the papers in this volume exploits the advent of diachronic treebanks to subject longstanding issues to large-scale data analysis, which was not possible before these resources. For example, Taylor & Pintzuk analyze split coordination in English. Crucially, PoS-tagged corpora would not suffice to allow data extraction to the same extent as a treebank does, and to collect a sufficient number of occurrences of the construction studied in this paper. Split coordination in English is not a construction that can be located automatically in either plain text or even simple PoS-tagged corpora, and both its long period of attestation and its relative rarity precludes manual searching. The treebank-based study by Taylor & Pintzuk reveals an interesting result that could not possibly have been reached without the help of this type of resource: "split coordination" in fact represents two different constructions, one of which remains stable over time while the other is lost in the post-Middle English period.

The volume is not restricted to studies of syntactic change: Eckhoff demonstrates how enriched treebank data can be employed in an analysis of an important semantic-morphological change in Russian: the rise of the *po* delimitative and its consequences for the Russian aspect system. The main focus is on the association between derivational morphology and semantics. To achieve this, verbs in the treebank have been enriched with tags indicating their internal structure (prefixation, stem, suffixation), as well as semantic tags. Thus, information about word-internal structure and semantics may be combined with the morphological and syntactic information that is already present in the treebank. The syntactic level makes it possible to identify potential delimitative contexts. Eckhoff employs all of this data to reevaluate the chronology of the change quite substantially.

Diachronic treebanks with good coverage also make large-scale statistic modeling possible. Ponti & Luraghi use treebank data to model a major syntactic shift: the rise of configurationality in Greek and Latin. Notably, data extraction from treebanks allows examining issues that could not easily be addressed based on PoS-tagged corpora, such as the decline of null anaphora for referential null objects. This feature was approximated by the absolute frequency of the part-of-speech tags of the nodes adjacent to verbs. The loss of zero anaphora is expected to be related with the skyrocketing of the rate of (personal) pronouns in that position for late varieties, which turns out to be the case. This paper is also a good example of another common feature of the papers mentioned above: the co-existence of questions (and related literature) of historical linguistics with methods for data analysis and evaluation borrowed from other disciplines. For example, to run network analysis on data in Ancient Greek and Latin, Ponti & Luraghi apply a tool developed in the field of computational biology both to build the networks and to calculate the topological indices used to evaluate their physical properties. The results are then interpreted linguistically in order to answer the question about the rise of configurationality in these languages.

Overall, addressing core linguistic questions, exploiting empirical evidence enhanced with metalinguistic information, applying multi-disciplinary techniques of data analysis and providing replicable results are all distinctive characteristics of the papers included in the volume.

5.2 Treebanks, text attestations and methodology

In historical linguistics, the available text sources are a perpetual problem. We are restricted to what written sources come down to us, and the inventory we are left with is normally skewed, to some extent random and often very sparse. The fact that we are restricted to only written sources is itself a limitation, since even poorly standardized written material tends to be conservative and will not necessarily reflect ongoing linguistic change. Some genres, such as religious texts, legal documents and metric poetry, tend to be more archaic than more narrative genres, often even formulaic. The advent of richly annotated diachronic treebanks provides new ways to handle the challenges posed by the state of the sources. Two of the papers in this volume make a direct methodological contribution to this problem, by applying treebank data and statistical modeling to assess the relationship between attested texts and the vernacular.

Simonenko, Crabbé & Prévost address the issue of genre differences, by looking at the relationship between prose and verse in historical French. There is a long-held intuition that prose is more progressive than verse in reflecting linguistic change. Statistically modeling two major syntactic changes in the history of French, the loss of null subjects and the loss of OV word order, the initial result is that there appears to be a significant difference between the rates of change in prose and verse. However, casting their analysis in terms of grammar competition and operationalizing the competing abstract grammars, the authors are able to show that the pace of change is the same in prose and verse if one corrects for metalinguistic factors. Their paper thus demonstrates both the analytic and methodological advances a large-scale treebank study can offer in combination with abstract grammatical representations.

In his study of late Latin charters, Korkiakangas complements the study by Simonenko, Crabbé & Prévost by using treebank data to assess the relationship between written text and formulaicity within single texts rather than between genres. Charters are legal documents and always contain a considerable share of formulaic language, but they also always contain a "free" part in which the case in question is described. Korkiakangas exploits this difference to measure how likely various types of change are to be visible in the formulaic part of the charters. He selects a number of phenomena that are commonly considered to be undergoing change in late Latin, and measures the extent to which they are reflected in the formulaic and free parts respectively. He demonstrates that a number of innovations are in fact found in both text types, and concludes that only particularly salient phenomena, either perceptually or syntactically, are preserved in their conservative form in the formulaic parts of the charters. Studies of this kind make it possible to make better use of conservative text genres in linguistic studies, rather than discarding them altogether.

6 Conclusions

The papers in this volume demonstrate the state-of-the-art use of diachronic treebanks in historical linguistics. They offer considerable advances, not only by providing structured data that allow innovative interpretations of long-standing issues, but also by opening new methodological avenues. A good treebank annotated according to the existing standards in the field enables the researcher to be explicit about her theoretical assumptions and selectional criteria, and opens for replicable studies.

Large-scale treebank data also allows filtering of problematic data in ways that were not available to us before.

There is still much work to do in treebanking for ancient languages, and it looks as though syntactic annotation is just a step on a longer path. Together with both building new treebanks for still under-resourced languages and enlarging the already available ones, the research community dealing with linguistic resources for ancient languages is now in the process of enhancing textual corpora with different layers of semantic and textual information, including semantic role labeling, ellipsis resolution and coreference analysis. This new trend is already being pursued by some of our authors (as e.g. in Eckhoff's paper), and hopefully such efforts will soon impact the world of diachronic studies in linguistics, supporting research across all linguistic levels.

References

Bamman, David & Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*,79–98. Berlin: Springer.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdt3.0/

Eckhoff Hanne M. & Aleksandrs Berdičevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15. 9–25.

Eckhoff, Hanne Martine, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Trygve Truslew Haug, Odd Einar Haugen & Marius Larsen Jøhndal. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52(1). 29–65.

Freddi, Maria & Silvia Luraghi. 2013. Appendix – Resources in syntax. In Silvia Luraghi & Claudia Parodi (eds.) *The Bloomsbury Companion to Syntax*, 463–468. London/New York: Bloomsbury.

Haug, Dag Trygve Truslew. 2015. Treebanks in historical linguistic research. In Carlotta Viti (ed.), *Perspectives on Historical Syntax*, 185–202. Amsterdam: John Benjamins.

Haug Dag Trygve Truslew & Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop* (LaTeCH 2008), 27–34. Marrakech, Morocco, 1st June.

Korkiakangas, Timo & Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal of Language Technology and Computational Linguistics* 26. 103–114.

Kroch Anthony, Beatrice Santorini & Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania.

Kroch Anthony & Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2).* Department of Linguistics, University of Pennsylvania.

Martineau, France. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus* 7. http://corpus.revues.org/1508

McEnery, Tony, Richard Xiao & Yuko Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Passarotti, Marco. 2011. Language Resources. The State of the Art of Latin and the Index Thomisticus Treebank Project. In *Deuxiéme Colloque International ALIENTO*, 301–320. ALIENTO.

Sinclair, John. 2005. Corpus and Text - Basic Principles. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books: 1–16. Available online from http://ota.ox.ac.uk/documents/creating/dlc/

Stein, Achim and Sophie Prévost. 2013. Syntactic annotation of medieval texts : The Syntactic Reference Corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible and Richard Whitt (eds.), *New Methods in Historical Corpora*, 75–82.Tübingen: Narr.

Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. *The York-Toronto-Helsinki parsed corpus of Old English Prose*. University of York.

Taylor, Ann, Marcus Mitchell & Beatrice Santorini. 2003. The Penn Treebank: An Overview. In Anne Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, 5–22. Dordrecht: Kluwer Academic Publishers.